McGill University

Desautels Faculty of Management

MGSC401: Statistical Foundations of Data Analytics

A Statistical Analysis of Modern Film: Understanding How to Achieve a High IMDb Score

Submitted to Professor Juan Camilo Serpa

Gianna Buenaventura
Grace Candiotti
Grace Francoeur
Youcef Sahnoune
Jonathan Steinberg
Vaughn Trestan

June 2nd, 2021

Contents

1 Introduction

Every year, hundreds of millions of dollars are poured into the film industry, and yet, successful films are far and few in between. For every one film that suspends an audience's disbelief and connects them through its story and characters, thousands of films remain golden ticket-less, never coming close to or even making the box office charts. Some films garner much pre-release buzz but flop with public audiences, whereas others are labelled as coming-of-age or cult classics despite flying under critics' radars. Consequently, film-making might seem like a box of chocolates (executives and audiences alike never know what they're going to get). But, what if there were a way to predict the success of an upcoming film?

Using various statistical modelling and data analysis techniques, this project will analyze characteristics of a wide variety of movies to understand the factors that position a movie for the highest levels of success. To quantify the success of a film, we will use IMDb, an online movie database that allows registered users to rate films. Because the database compiles a weighted average of thousands of user ratings to determine a movie's overall score, the metric provides visibility into a film's popularity and likability among the public. And for many, after a long day of work or school, an IMDb score is the last line of defense between what could be an unforgettable movie night and well, the Emoji Movie.

As it stands, for members of the film industry including directors, producers, movie executives, and actors alike, understanding the factors that contribute to a high IMDb score is extremely valuable. From marketing spend to consumer research to actual production, understanding how to leverage the free, word-of-mouth promotion IMDb can provide, will be critical in directing their decisions when budgeting for future productions and films. With considerations like the genre of a film and a film's optimal duration, our statistical model aims to accurately predict how successful a film will be, ideally providing insight for upcoming chef d'oeuvres.

2 Data Description

In this study, the data we've utilized comes from IMDb, an online database of information with a focus on films and television shows. Here, we'll be analyzing a set of 3,987 movies filmed over the last 80 years. With each observation, we look at twenty quantitative and qualitative dimensions, ranging from production year to movie genre to the lead actor's popularity, to build a predictive model (see Table 1). Here, we'll introduce the methodology we used to understand each explanatory variable, and outline how each is related to our dependent variable, IMDb score.

2.1 Quantitative Variables

When analyzing our quantitative variables, we started by exploring their distributions. Our dependent variable, *IMDb Score*, was relatively normally distributed, with a bit of negative skewness (Appendix, Figure 1).The average score among the selected films was 6.46 out of 10, with a standard deviation of 1.085. Additionally, the range was relatively well spread out, with ratings ranging from 1.6 (Justin Bieber really messed up on this one) to 9.3. Moreover, Referring to Figures 2 and 3 in the appendix, we illustrated each of our explanatory variable's distributions both with histograms and box plots. These models were integral in understanding the unique qualities of each variable, and were a first start at examining the regression power (with respect to linearity) of each. Throughout the dataset however, we saw a significant amount of skewness in the predictors. Of the 10 quantitative predictors, all but *Film Year* were skewed to the right. This analysis helped us garner some valuable insights, as we saw that nearly 87% of the movies were made in the last 30 years, the average movie length was just under 2 hours, and that nearly 46% of movies didn't have a single Facebook like. These findings implied

Variable List	
IMDb Score	*A crowdsourced movie rating, out of 10.*
Title Year	*The year the movie was released.*
Movie Duration	*The length of the film, in minutes*
Main Genre	*The movie's primary genre.*
Secondary Genre	*The movie's secondary genre.*
Plot Keywords	*Key aspects of the movie's plot.*
Language	*The language the movie was filmed in.*
Country	*The country the movied was released in.*
Content Rating	*The film's content rating.*
Budget	*The film's budget, in local currency.*
Aspect Ratio	*The film's presented aspect ratio.*
Number of Faces	*The # of faces on the film's poster.*
Movie's Likes	*The # of Facebook likes on the movie's page.*
Actor 1 Name	*The lead actor's first and last name.*
Actor 2 Name	*The secondary actor's first and last name.*
Actor 3 Name	*The third actor's first and last name.*
Actor 1 Likes	*The # of Facebook likes on actor one's page.*
Actor 2 Likes	*The # of Facebook likes on actor two's page.*
Actor 3 Likes	*The # of Facebook likes on actor three's page.*
Director Likes	*The # of Facebook likes on the director's page.*

Table 1: Data Dictionary

quite nonuniform data throughout the set that we'd have to keep in mind when making any managerial recommendations, with high positive skewness driven by an increased frequency of very low bounds in the data. For example, as mentioned with Facebook likes, the high number of actors or directors with zero likes skewed the distributions significantly.

We then dove deeper into the data, examining if the variables had any errant observations or outliers that could have an impact on our later models. Based on our histograms and box plots, it was clear to see that visually, outliers were present (Appendix, Figures 2 and 3). When looking to actually examine which variables had significant outliers, we landed on both *Actor 1 and Actor 2 Likes*, *Movie Likes*, the movie's total *Budget* (in local currency), and *Aspect Ratio*. Our criteria here was simple: we looked for observations that affected how our function of choice fit the data, and we focused on predictors that had a material impact. Logically, it was clear as to why these variables in particular would have some unruly observations. When it comes to actor likes, there are significant gulfs in popularity for different actors and actresses, and this is reflected in social media. Similarly, depending on the quality of the cast and marketing spend, the likes a movie can get can also vary widely. Additionally, due to the fact that movie budgets were expressed in local currency, it was no surprise that there were some numbers of insane magnitude (solely due to the fact that they were funded in a non-USD currency). Lastly, we noticed that there were some extremely high and unreasonable aspect ratio values in the dataset, but we attributed these mostly to mistakes in data collection.

Moreover, after understanding the nuances associated with our predictors and their distributions, we wanted to check whether or not collinearity was present throughout the variables. After running a correlation matrix across each of the quantitative variables (Appendix, Figure 5), we found that there wasn't any significant correlation between our variables as none of the variables were highly correlated (correlation$>$0.8). It would be beneficial to note however, that there were mild amounts of correlation between actor likes on Facebook, which stands to reason as many movies will typically have lead actors and actresses of the same or similar reputation/popularity. In general, it was valuable for us to see that collinearity wouldn't impact our regressions, at least among our quantitative variables.

Following our analysis of the variables and their shape, any outliers, and any potential issues with

collinearity, we checked for any linear relationships between the data and *IMDb Score*. Importantly, we found that linear predictive power across the board was generally lacking. Only two variables, the movie's duration, and the movie's Facebook likes, had r-squared values above 0.04, (Appendix, Table 2) with every other variable sorely lacking on predictive power. However, with regard to p-values, each variable outside of aspect ratio and movie budget did show significance, which warranted further investigation. In Section 3.1 of the paper, we will outline additional tests we conducted to understand the relationships between the set's quantitative predictors and *IMDb Score*.

2.2 Qualitative Variables

In addition to the aforementioned quantitative predictors, we explored ten qualitative independent variables to better understand our dataset. Of these ten, five predictors had few enough categories to allow for comment on their distribution. Notably, over 78% of films were created in the USA, over 95% were filmed in English (with English being one of 34 languages), and 45% of movies had an advisory rating of R.

Contrasting the three previous variables, the main and secondary genres of a movie demonstrated more segmented distributions. The two biggest *Main Genre* categories of Comedy and Action comprised approximately half of observations (26% and 25%, respectively), leaving 17 other genre possibilities for the remaining 49% of films. As for the *Secondary Genre* of a movie, Dramas and Adventure films encompassed the most observations (24% and 10%, respectively), with 20 other genres splitting the remaining observations. The variables for *Actor Name 1, 2, and 3*, *Director Name*, and *Plot Keywords* demonstrated no significant distribution frequency due to each variable's many unique possible categories.

To examine the relationship between *IMDb Score* and each qualitative variable, we created simple linear regressions to understand a given variable's predictive power and statistical significance. Because *Country* and *Language* were clearly segmented, we created two categorical variables: *USA* and *English*. These variables denoted whether each observation was filmed in the US and whether it was filmed in English, respectively. As for *Plot Keywords*, this variable demonstrated a very high adjusted r-squared of 90%. However, because there were so many unique values, we attempted to segment it by analyzing the most frequently used plot keywords. But, this segmentation ended up being redundant because the categorizations were akin to those for *Main* and *Secondary Genre*. In general, for the majority of qualitative variables, adjusted r-squared values were relatively low. For variables like *Actor 1, 2, and 3*, this was understandable given the large number of unique categories for each variable. Like the quantitative predictors, many qualitative variables had low p-values denoting statistical significance, which led to further inquiry during our model-building process.

3 The Model Selection

3.1 Variable Selection

The methodology for building our model began by using the aforementioned simple linear regressions at least on the quantitative side to determine the variables that demonstrated statistical significance and relatively high predictive capabilities (Appendix, Table 2). This was completed by analyzing each coefficient's p-value and the model's adjusted r-squared value. After completing this, we moved away from the assumption of linearity and ran ANOVA tests for each quantitative variable to determine the polynomial degree that optimally fit each variable, and then evaluated their relationship to IMDb Score.

Quantitative Predictors: As shown by the regressions, linear predictive power of the variables were generally lacking. There were some notable standouts, however, such as *Movie Duration* and *Movie Likes*,

that had relatively higher r-squared values (Appendix, Table 2). As mentioned, we also built ANOVA models to fit the other variables, in the case that linearity was the only reason that predictive power was low. However, we did note that significance did not materially differ between linear and ideal fit for the bulk of our variables (outside of the aforementioned *Movie Duration* and *Movie Likes*). It is also important to note that there were general logical reasons behind omitting some predictors. For example, Budget contained a significant number of differing currencies and values, which were not standardized and currency- or inflation-adjusted across the 80 years of films; this made it hard for us to drive any meaningful managerial insights.

Qualitative Predictors: To further explore the qualitative predictors, we investigated each of the regressions' significance and predictive power. Despite statistical significance for both binary variables *USA* and *English*, the variables' lack of predictive power (r-squared of 1.36% and 2.73%) resulted in our exclusion of either variable in our model. Similarly, the regressions for each *Actor Name* variable (the first, second, and third named actors in a given movie) and *Director Name* resulted in exclusion of the five variables. Although the relationships demonstrated high predictive power, almost every variable was statistically insignificant and the model's adjusted r-squared was low. Consequently, we attempted to segment the *Actor* variables into different categories (A-list, B-list, C-list) which depended on the number of Facebook likes a given actor had on their page (which we assumed was related to actor talent). Even with this categorization, the variable's predictive power remained low, and hence, we excluded it when building our model. After analyzing the relationship between *IMDb Score* and the remaining qualitative variables, Main Genre was the only qualitative predictor used in our model, as its simple regression displayed the highest predictive power (Appendix, Table 7).

As such, our analysis identified *Main Genre, Movie Duration*, and *Movie Likes* as the best variables to include in our model. *Main Genre* was fitted as a categorical variable with 17 dummy variables and yielded an adjusted r-squared score of 10.8% with many statistically significant coefficients (Appendix, Table 7). Next, an ANOVA test determined that *Movie Duration* achieved its best fit as a polynomial of degree five, which resulted in an adjusted r-squared score of 16.31% (Appendix, Table 3). Finally, *Movie Likes* was fitted using a polynomial spline regression, which produced an adjusted r-squared score of 17.13%. By visualizing the scatter plot and conducting an ANOVA test, we determined that three knots at 1,000, 10,000, and 50,000 likes with a polynomial regression of degree nine resulted in the best fit (Appendix, Table 4). However, because the p-value associated with degree nine was close to 0.05, we opted to use degree eight instead due to its lower p-value. The combination of these three variables gave us our first complete model as seen below in *Equation 1*, which has an overall adjusted r-squared value of 32.5%.

$$\hat{score} = genre_{\text{categorical}} + duration_{\text{d=5}} + movie\_likes_{\text{3 knots, d=8}} \tag{1}$$

3.2 Generating Votes Variable and Model 2

In an attempt to improve our model's predictive capabilities, we reexamined the dataset and found that the *Movie Link* label gave us access to more valuable information. Specifically, because IMDB scores are crowdsourced, we hypothesized that creating a variable called *Votes*, which denoted the number of votes per movie on IMDb, could increase our model's predictive power. Thus, we scraped the number of votes on IMDb from each movie through *Movie Link* and added an additional column to our dataset. As performed with all other variables, an ANOVA test determined that a polynomial regression of degree five resulted in the best fit for Votes (Appendix, Table 5). The regression's adjusted r-squared value of 21% suggested that this variable could benefit our model.

To determine the combination of variables with the highest predictive power, we ran a series of regressions by adding one variable at a time then assessing its effect on adjusted r-squared. The regression with just *Votes* and *Main Genre* resulted in an adjusted r-squared value of approximately 38%, but adding *Movie Likes* and *Movie Duration* led to a higher adjusted r-squared of 41.48%. Thus, we included all four variables in our second model, as seen below in *Equation 2*.

$$\hat{score} = genre_{\text{categorical}} + duration_{\text{d=5}} + movie\_likes_{\text{3 knots, d=8}} + votes_{\text{d=5}} \tag{2}$$

3.3 Cross Validation: The K-Fold Test

Following the creation of Model 2, we noted that including four variables only marginally increased adjusted r-squared and thus increased the risk of overfitting our model to the training dataset. In comparison to the aforementioned model containing only *Votes* and *Main Genre*, Model 2's adjusted r-squared increased by only 3% (whereas the number of variables doubled). To determine whether overfitting was present in Model 2, we performed K-fold tests to identify which model produced a lower mean squared error (MSE). We chose to use K-fold testing because of its more reliable and less variable MSE values than Validation Set testing and its speed advantage over LOOCV tests. Our choice of 20 folds ensured that our output produced less variable results while running in a reasonable amount of time. To perform K-fold testing, we grouped the least frequently observed *Main Genre* categories into one single category to reduce the possibility of errors in our testing. Seven genres were observed to have approximately 20 observations or fewer and were grouped into a category called *"Other."* As a result, we were left with 11 dummy variables in the *Main Genre* categorical variable. This slightly decreased the adjusted r-squared values for both models but was necessary in order to perform the resampling testing.

Surprising as it was, the K-fold test for Model 2 resulted in an MSE value of greater than 1000. Consequently, we performed a Variance Inflation Factor test and determined that there was collinearity amongst polynomial degrees greater than two for *Movie Likes* (Appendix, Table 7). To eliminate the collinearity, we altered the degree of Movie Facebook Likes to a spline regression with a degree of two. After making this adjustment, the updated K-fold test for Model 2 produced an MSE of 0.69. This MSE value was lower than that of the two-variable model and suggested an absence of overfitting in Model 2. After adjusting the degree of *Movie Likes*, we created a third and final model, which can be seen below in *Equation 3*. Through performing an F-test to validate whether or not this model was strong, we concluded that this overall model is statistically significant and possesses an adjusted r-squared value of 41.15%.

$$\hat{score} = genre_{\text{categorical}} + duration_{\text{d=5}} + movie\_likes_{\text{3 knots, d=2}} + votes_{\text{d=5}} \tag{3}$$

To determine whether there was any significant interaction between the predictors in our final model, we ran regressions to detect interaction. Specifically, we tested all combinations of our predictor variables with an interaction term, but in the end, we concluded that there was no significant interaction in our model based on the statistically insignificant p-values of our predictor interaction terms. Lastly, we tested for the presence of heteroskedasticity in two ways. We first ran a plot of our model's residuals to observe its shape (i.e. whether the plot had non-constant variance and was therefore shaped like a funnel). In spite of the test's encouraging results, we then performed a formal non-constant variance (NCV) test to confirm our suspicions. The test resulted in a p-value greater than 0.05, validating the heteroskedasticity in Model 3, and as such, we performed a regression which corrected for the heteroskedasticity. The correction did not alter our model's coefficients but resulted in the decrease of some p-values of our predictors. However, as the corrected model did not produce an r-squared result, we concluded that our model possesses the same level of predictability of 41.15%, as was previously found.

4 Managerial Implications

Through an analysis of our final model, we can conclude that *Movie Genre, Movie Duration*, and *Movie Likes*, and *Votes* for ratings on IMDb impacts a given movie's IMDb score. The model's coefficients show the individual effects of each independent variable on the dependent variable, being the IMDb score. We found that of the quantitative variables, *Movie Likes* and *Votes* had a positive relationship to *IMDb Score*, while *Movie Duration* was negative. As for *Movie Genre*, the results were varied, and we found that some genres are correlated to a higher predicted score than others.

We concluded that shorter movies tended to correlate with higher IMDb scores (while still noting that there remains outliers for this statement). As such, production teams should aim for shorter movies while ensuring that no integral aspects of a movie are left out. Thus, during every step of the movie-making process, it is imperative that production teams work closely with the creative teams to ensure that there are no miscommunications or silos between the aspects of the film to optimize its length. As important as it is for directors to obtain creative reign in their decisions on the scenes to keep and cut from a film, it is also important that most, if not all, scenes resonate with audiences to maximize their likelihood of favourably rating.

As for the genre of a film, most categories result in increased IMDb scores. But, there are a few genres that negatively impact an IMDb score, which is a significant consideration to make as a movie executive, writer, director, or otherwise. Specifically, action and horror movies result in the lowest predicted score in relation to the other aforementioned genres used in our model. In terms of horror movies, the lower scores may be explained by the fact that many horror movies are not suitable for select audiences due to graphic visuals, with subject matter potentially breaching points of contention or controversy (and thus, negatively affecting IMDb score). Conversely, documentaries are predicted to have the highest amount of success over other genres. Just as societal culture and norms evolve over time, societal movie preferences and trends evolve, too. As was previously alluded to, it's important for decision makers in the film industry to be in tune with audience preferences as the public will likely dictate the trajectory of a film's success.

As a marketing or promotion executive in the film industry, understanding how to increase a movie's Facebook likes will optimize a movie's chance at success. Because the predictive variable is positively correlated to a film's IMDb score (i.e. more likes relates to higher scores), marketing, promotions, and public relations teams should focus on garnering the most visibility and awareness for a movie on Facebook. In the very least, films should create and maintain Facebook pages in addition to posting promotional content about the movie's trailer, website, and upcoming events. Additionally, engaging in social media trends and hosting contests and giveaways could allow for certain movie pages to "trend" on Facebook due to increased engagement with audiences. Notably, experiential marketing has gained popularity in recent years due to its ability to transport audiences to and connect audiences with a film's world. Additionally, it has become easier for audiences around the world to partake in experiential marketing because of digital marketing. By promoting this and other types of content on Facebook, a given movie will expand its online presence, and thus, increase its likelihood of amassing Facebook likes and therefore, a movie's IMDb Score.

Finally, the number of voters on IMDb that a given movie receives is positively correlated to IMDb score. As the number of voters increases, the weighted average (that is, the IMDb score) encompasses more data. If a movie gets a poor review, the weight of that user's score is smaller if there are more votes to weigh it against. Contrarily, if a movie has fewer voters, each individual's opinion carries more weight, and will thus have more impact on an IMDb score. Because the types of movies that receive more votes are likely those that either extremely positively or negatively affect audiences, it makes intuitive sense for votes to positively correlate with IMDb score. To increase a film's IMDb vote count, marketing

and promotions teams can utilize similar recommendations as those to increase a film's Facebook likes due to the online-based nature of the metric. As the public gains more exposure to the movie through promotions, the more likely it will be that the movie receives more votes. In this case, executives should try to stray from undertaking film projects with plotlines and characters that will leave audiences feeling indifferent post-view. Because this is not always an easy feat, production companies may want to consider conducting field research on the types of movies that audiences would be the most interested in and resonate the most with.

Based on our model, directors and producers should devote more time to the logistics of a movie like duration while still taking subjective factors into consideration. Notably, the two factors which garnered the most influence on a movie's IMDb score are (in this order): 1) genre, and 2) the number of IMDb votes a movie receives. Although a strict ruling on deciding the genre of a movie may turn some directors away from a project, the takeaway from this outcome is that there are proactive measures that production companies can take to position a film for the most success.

At the end of the day, whether you're at home or in the theater, one thing remains clear: there's no perfect movie (outside of Cars 2 (2011), of course), and statistics is not an exact science. But, after a robust analysis of roughly 4,000 films, we can tell you that with a well selected genre, movie length, and adequate marketing, you might just get close. And beyond that, in the words of Keanu Reeves, "where we go from there is a choice I leave to you."

5 Appendix

Figure 1: **Dependent Variable: IMDb Score**

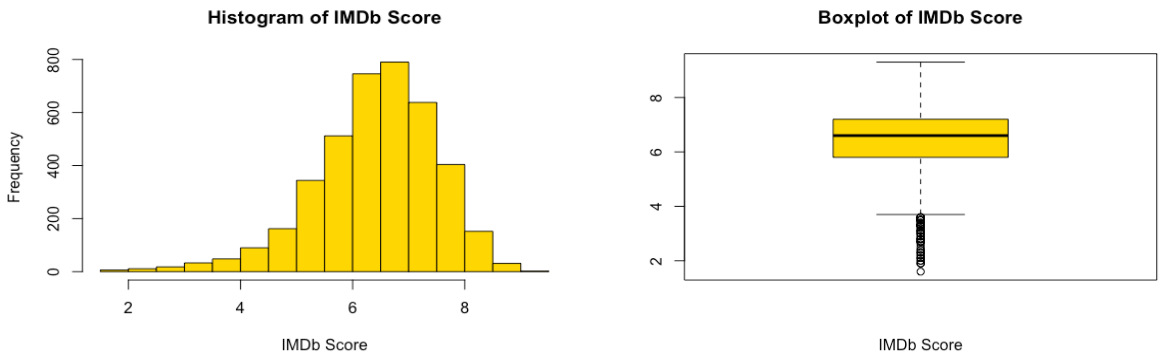


Figure 2: **Quantitative Predictors: Histograms**

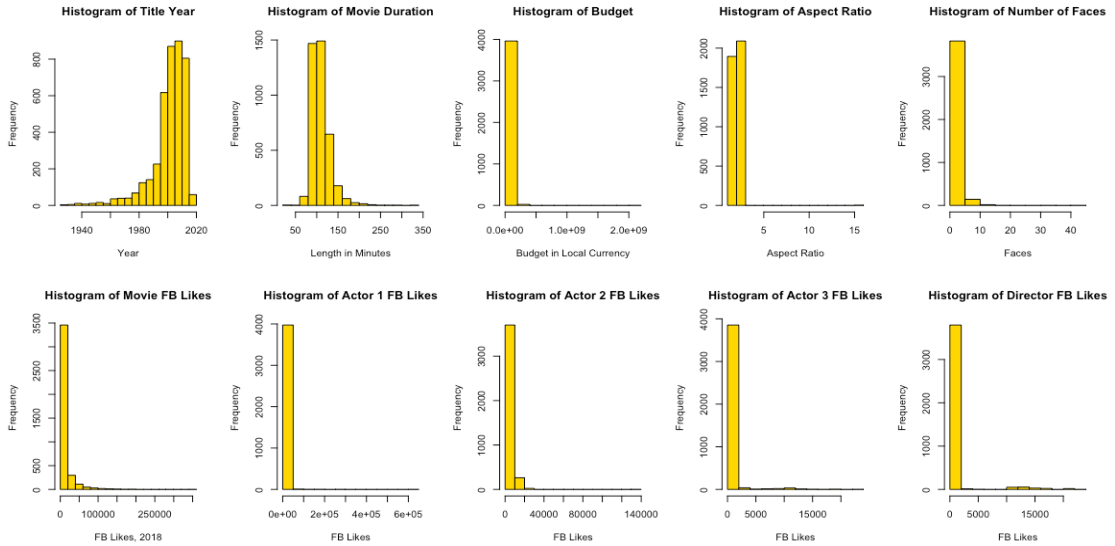


Figure 3: **Quantitative Predictors: Box Plots**

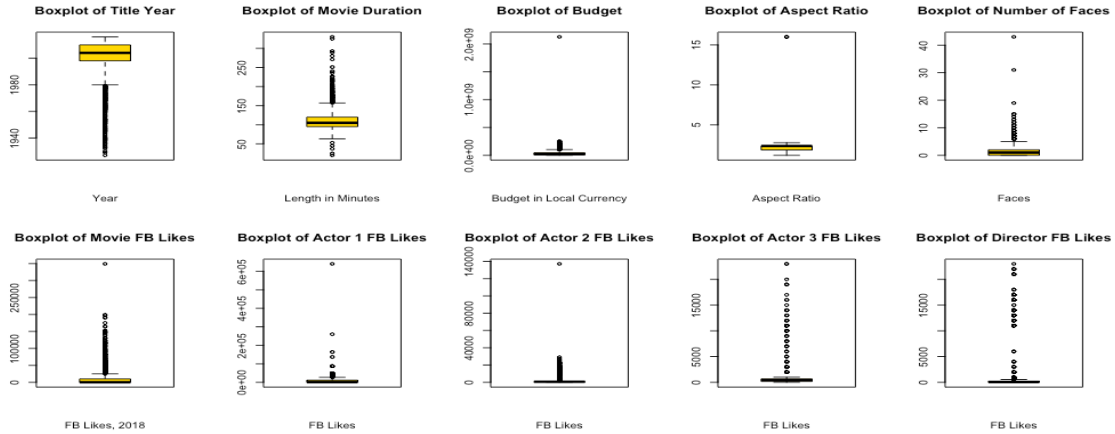


Figure 4: **Qualitative Predictors: Relative Frequency**

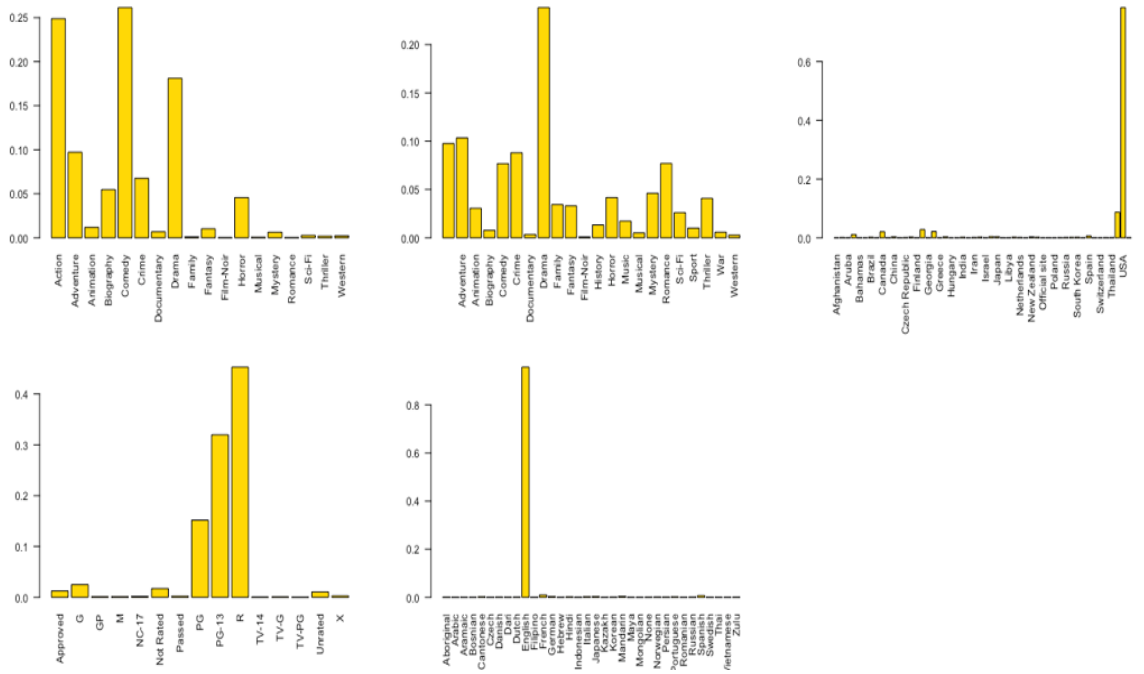
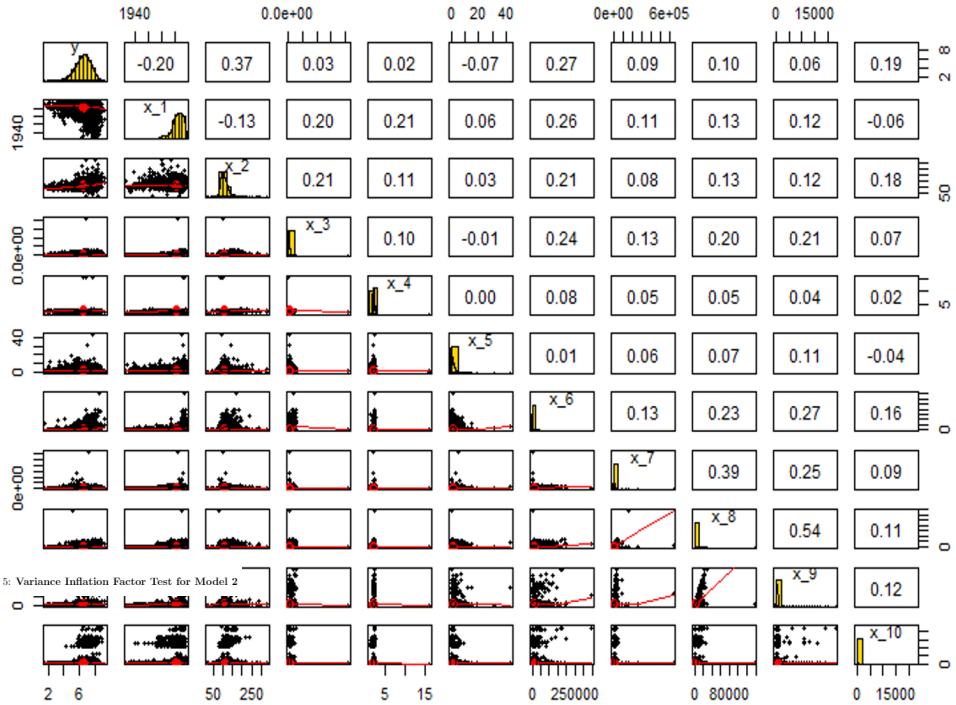


Figure 5: **Correlation Matrix**



y - IMDb Score
x_1 - Title Year
x_2 - Movie Duration
x_3 - Budget
x_4 - Aspect Ratio
x_5 - # of Faces
x_6 - Movie Likes
x_7 - Actor 1 Likes
x_8 - Actor 2 Likes
x_9 - Actor 3 Likes
x_{10} - Director Likes

Table 5: **Variance Inflation Factor Test for Model 2**

Table 2: **Simple Regressions: Quantitative Predictors**

	Dependent variable:									
	IMDb Score									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Title Year	−0.017***									
	(0.001)									
Movie Duration		0.018***								
		(0.001)								
Movie Budget			0.000*							
			(0.000)							
Number of Faces on Poster				−0.035***						
				(0.008)						
Aspect Ratio					0.040					
					(0.036)					
Movie Likes						0.00001***				
						(0.00000)				
Actor 1 Likes							0.00001***			
							(0.00000)			
Actor 2 Likes								0.00002***		
								(0.00000)		
Actor 3 Likes									0.00004***	
									(0.00001)	
Director Likes										0.0001***
										(0.00001)
Constant	41.133***	4.495***	6.435***	6.504***	6.372***	6.334***	6.411***	6.412***	6.429***	6.403***
	(2.720)	(0.079)	(0.021)	(0.021)	(0.078)	(0.018)	(0.019)	(0.019)	(0.018)	(0.017)
Observations	3,987	3,987	3,987	3,987	3,987	3,987	3,987	3,987	3,987	3,987
R^2	0.039	0.139	0.001	0.004	0.0003	0.075	0.008	0.009	0.004	0.036
Adjusted R^2	0.039	0.138	0.001	0.004	0.0001	0.075	0.008	0.009	0.004	0.035
Residual Std. Error (df = 3985)	1.064	1.008	1.085	1.083	1.085	1.044	1.081	1.081	1.083	1.066
F Statistic (df = 1; 3985)	162.531***	641.096***	3.370*	17.145***	1.224	321.986***	31.235***	36.663***	16.403***	147.504***

Note: *p<0.1; **p<0.05; ***p<0.01

Table 3: **ANOVA: IMDB Score x Movie Duration, Polynomial, Degrees 1-6**

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	3985	4046.03				
2	3984	3993.98	1	52.05	52.79	0.0000
3	3983	3983.46	1	10.52	10.67	0.0011
4	3982	3937.91	1	45.55	46.19	0.0000
5	3981	3925.95	1	11.96	12.13	0.0005
6	3980	3924.43	1	1.52	1.54	0.2148

Table 4: **ANOVA: IMDB Score x Movie Likes, Spline, Degrees 1-10**

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	3982	4039.19				
2	3981	3968.92	1	70.28	72.05	0.0000
3	3980	3933.32	1	35.60	36.49	0.0000
4	3979	3913.75	1	19.57	20.07	0.0000
5	3978	3900.30	1	13.45	13.79	0.0002
6	3977	3892.81	1	7.48	7.67	0.0056
7	3976	3887.18	1	5.63	5.77	0.0163
8	3975	3881.70	1	5.48	5.62	0.0178
9	3974	3877.84	1	3.87	3.97	0.0465
10	3973	3875.10	1	2.74	2.81	0.0940

Table 5: **ANOVA: IMDB Score x Movie Votes, Polynomial, Degrees 1-10**

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	3985	3668.63				
2	3984	3520.76	1	147.87	169.85	0.0000
3	3983	3484.05	1	36.71	42.16	0.0000
4	3982	3471.59	1	12.46	14.31	0.0002
5	3981	3465.05	1	6.54	7.52	0.0061
6	3980	3462.93	1	2.11	2.43	0.1194
7	3979	3461.90	1	1.03	1.18	0.2765
8	3978	3461.87	1	0.03	0.04	0.8493
9	3977	3461.75	1	0.12	0.14	0.7096
10	3976	3461.56	1	0.19	0.22	0.6381

Table 6: **Variance Inflation Test for Model 2**

	Variables	**Tolerance**	**VIF**
1	Main Genre: Adventure	7.713415e-01	1.296443
2	Main Genre: Animation	9.220260e-01	1.084568
3	Main Genre: Biography	8.202643e-01	1.219119
4	Main Genre: Comedy	6.168253e-01	1.621204
5	Main Genre: Crime	8.315514e-01	1.202571
6	Main Genre: Documentary	9.488046e-01	1.053958
7	Main Genre: Drama	6.649677e-01	1.503833
8	Main Genre: Fantasy	9.611035e-01	1.040471
9	Main Genre: Horror	8.535445e-01	1.171585
10	Main Genre: Mystery	9.766640e-01	1.023894
11	Main Genre: Other	9.684145e-01	1.032616
12	Movie Duration, degree 5, 1	7.616469e-01	1.312944
13	Movie Duration, degree 5, 2	9.294218e-01	1.075938
14	Movie Duration, degree 5, 3	9.476550e-01	1.055236
15	Movie Duration, degree 5, 4	9.780285e-01	1.022465
16	Movie Duration, degree 5, 5	9.744134e-01	1.026258
17	Spline, Movie Likes, degree 8, 1	4.92425e-01	2.329686
18	Spline, Movie Likes, degree 8, 2	4.642865e-01	2.143843
19	Spline, Movie Likes, degree 8, 3	4.945785e-02	20.219236
20	Spline, Movie Likes, degree 8, 4	4.216126e-03	237.184587
21	Spline, Movie Likes, degree 8, 5	4.999181e-04	2000.327530
22	Spline, Movie Likes, degree 8, 6	8.469163e-05	11807.541934
23	Spline, Movie Likes, degree 8, 7	2.333884e-05	42847.031708
24	Spline, Movie Likes, degree 8, 8	1.256100e-05	79611.468751
25	Spline, Movie Likes, degree 8, 9	1.852714e-05	53974.863484
26	Spline, Movie Likes, degree 8, 10	1.520583e-04	6576.425406
27	Spline, Movie Likes, degree 8, 11	9.439136e-01	1.059419
28	Total Votes, degree 5, 1	5.182984e-01	1.929391
29	Total Votes, degree 5, 2	7.649055e-01	1.307351
30	Total Votes, degree 5, 3	8.380474e-01	1.193250
31	Total Votes, degree 5, 4	8.325866e-01	1.201076
32	Total Votes, degree 5, 5	8.515485e-01	1.174331

Table 7: **Simple Regression: Movie Genre**

	Dependent variable:
	imdb_score
Adventure	0.306***
	(0.061)
Animation	0.456***
	(0.153)
Biography	0.938***
	(0.077)
Comedy	−0.081*
	(0.045)
Crime	0.644***
	(0.070)
Documentary	0.553***
	(0.200)
Drama	0.595***
	(0.050)
Family	0.571
	(0.514)
Fantasy	0.070
	(0.163)
Film-Noir	1.346
	(1.026)
Horror	−0.542***
	(0.083)
Musical	0.496
	(0.726)
Mystery	0.450**
	(0.208)
Romance	0.846
	(1.026)
Sci-Fi	0.009
	(0.311)
Thriller	0.331
	(0.389)
Western	0.512
	(0.343)
Constant	6.254***
	(0.033)
Observations	3,987
R^2	0.111
Adjusted R^2	0.108
Residual Std. Error	1.025 (df = 3969)
F Statistic	29.275*** (df = 17; 3969)
Note:	*p<0.1; **p<0.05; ***p<0.01